MODEST RISK SINGLE NUCLEOTIDE POLYMORPHISM POLYGENE IN BREAST CANCER SUSCEPTIBILITY

by

Tiana Christine Francy

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Oncological Science

The University of Utah

August 2013

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Tiana Christine Francy**

has been approved by the following supervisory committee members:

Sean Tavtigian	, Chair	**6/12/13** Date Approved
Deb Neklason	, Member	**6/13/13** Date Approved
Nicola Camp	, Member	**6/12/13** Date Approved

and by **Bradley Cairns** , Chair of

the Department of **Oncological Science**

and by Donna M. White, Interim Dean of The Graduate School.

ABSTRACT

During the 1990s, the search for breast cancer susceptibility genes identified several high-risk genes such as BRCA1&2, TP53, and PTEN. The late 1990s and early 2000s saw identification of the first intermediate risk genes, ATM, CHEK2, and PALB2. With time, however, the likelihood of identifying new high-risk genes is diminishing; on the other hand, the technology for hypothesis free, genome-wide identification of intermediate risk genes is still being developed. During the mid-2000s, the hunt for the genetic basis of breast cancer shifted toward genome-wide searches for modest risk breast cancer associated single nucleotide polymorphisms (SNPs). Here, we study a panel of 16 such SNPs as a polygene. We demonstrated that the polygene model works well for this panel, and that when viewed as a polygene, the composite risk for some individuals rises to a level comparable to an intermediate risk gene. We also find that the polygene model seems to better differentiate case-control status with increasing numbers of SNPs; therefore, we expect that the more SNPs that are added to the polygene model, the more accurately risk will be differentiated in the composite odds ratio (OR) and the closer the model will move toward medically actionable utility.

CONTENTS

ACKNOWLEDGEMENTS

I would like to thank Fabienne Lesueur, Florence Le Calvez-Kelm, Catherine Voegele, Nathalie Forey, and Nivo Robinot at IARC for preparing the whole genome amplified DNAs that I used for this genotyping, and for genotyping the markers FGFR2, TOX3, Map3K1, 8q24, LSP1, Casp8, and 2q35. I would like to thank the DNA Sequencing Core Facility, a part of the Health Sciences Cores at the University of Utah, for their work with genotype verifications. I would also like to thank Mikeal Wall of Biofire Diagnostics (previously Idaho Technologies) for his help with primer and probe design, and troubleshooting of the LightScanner. I would like to thank Alex Stark and Megan Evans for helping with marker development and production genotyping. I would like to thank Judith Rosenthal for day-to-day lab management. I'd like to thank Tonya DiSera for database help. The overall project was designed by Sean Tavtigian and funded by NIH R01 CA121245.

INTRODUCTION

Common Disease/Common Variant vs. Common Disease/Rare Variant

Historically, there have been two models to explain genetic predisposition in breast cancer: the common disease/common variant (CD/CV) model and the common disease/rare variant (CD/RV) model (Pritchard & Cox, 2001; Ivengar, 2007; Schork, 2009). The common disease/common variant model says that a limited set of sequence variants that are common in the population underlie the genetic basis of a common disease – in this case, breast cancer (Lander, 1996; Chakravarti, 1999; Reich, 2001). The second model argues that there are many rare variants in the population that underlie the genetic basis of a common disease (Pritchard, 2001).

Single Nucleotide Polymorphisms

With the increasing availability of human genome data, it has become clear that the differences from one human genome to the next often come down to single nucleotide base changes, termed as Single Nucleotide Polymorphisms (SNPs) (Chakravarti, 2001). Many SNPs have been identified through Genome-Wide Association Studies (GWAS) – studies that examine common variants in large numbers of individuals in order to determine whether any association exists between a particular variant and a specific trait (Manolio, 2010). It is important to note that in such cases, the SNPs which are found to be associated with a trait are often merely markers for genomic locations and are selected for their ability to adequately tag the genome (Easton and Eeles,

2008). That is, even if the SNP itself is noncoding, it could be tagging an underlying coding variant.

As is the case with a number of genetic diseases, many SNPs have been shown to be associated with breast cancer (Houlston, 2004). It is also known that SNPs associated with breast cancer risk are distributed throughout the genome, rather than being clustered in one particular area (Pharoah, 2002; Easton, 2007; Michailidou, 2013). The majority of SNPs associated with breast cancer are noncoding and are therefore thought to be regulatory (Easton, 2007).

The Polygene Model

Until early 2013, there were approximately eighteen well-established breast cancer risk SNPs (Wacholder, 2010). In March of 2013, a paper was published which established 41 new breast cancer susceptibility SNPs based on multiple Genome-Wide Association Studies (GWAS) (Michailidou, 2013). This study by the Collaborative Oncological Gene-environment Study (COGS), using a custom Illumina iSelect assay termed iCOGS, brought the total number of recognized breast cancer risk SNPs to at least 60, with possibly as many as 70 (Michailidou, 2013). Using the data from nine previous GWAS using individuals of primarily European descent, Michailidou et al. identified a panel of potential SNPs that could be shown to have an association with breast cancer. The authors then compiled samples from 52 previous studies that had participated in the Breast Cancer Association Consortium (BCAC) and designed a custom Illumina assay (iCOGS) with which to interrogate the potential new SNPs. When the genotyping of the samples was complete, the authors had identified over 40 new associations between SNPs and risk of breast cancer.

In terms of risk, the general female population is considered to have approximately a 10% lifetime risk of developing breast cancer (Howlader, 2009). Hence, a 10% lifetime risk roughly corresponds to an odds ratio of 1.0 on a logarithmic scale. A woman is considered to have a medically actionable level of genetic risk when her lifetime risk reaches approximately 20-25%, which roughly corresponds to an odds ratio of 2.0-2.5 (Harris, 2007; Ellsworth, 2010). Common breast cancer risk SNPs typically have odds ratios on the order of 1.01 to 1.3. Hence, the risk conferred by individual risk SNPs has been considered to be medically negligible (Pharoah, 2002; Easton, 2007; Michailidou, 2013).

The polygene model argues that while the increased level of risk is not additive across SNPs, it also cannot be discounted (Pharoah, 2002). The individual risk conferred by each SNP can be combined multiplicatively in order to determine a composite risk level conveyed by the unique combination of SNPs in each woman's genome. With many common, modest-risk alleles, the composite risk may climb to a medically actionable level. Additionally, with the very recent addition of a large number of SNPs newly identified as being associated with breast cancer by the COGS group, the potential for the discovery of polygenic effects of breast cancer risk SNPs has only increased.

Study Question and Hypotheses

The parent study for this project asked, "What is the relative contribution of common (usually modest-risk) sequence variants vs. rare (potentially higher-risk) sequence variants to the genetic population attributable fraction of breast cancer?" From that question arose our hypotheses. The null hypothesis for this project was that our empirical measure of odds ratio associated with the polygene would show no risk

associated with the polygene. Conversely, our alternate hypothesis was that the empirically measured risk associated with the polygene would be similar to a naïve product of all of a subject's risk SNP genotypes. In the event the null hypothesis is rejected, we wanted to estimate the fraction of the population, and the fraction of breast cancer cases, whose polygene risk is predicted to be medically actionable.

MATERIALS AND METHODS

Ethics Statement

The polygene genotyping studies and analysis described here were approved by the institutional review board (IRB) of the University of Utah, the International Agency for Research on Cancer IRB, and the local IRBs of the Breast Cancer Family Registry (Breast CFR) centers from which we received samples. All participants gave written, informed consent.

Subjects

Breast CFRs at three centers (Cancer Care Ontario, the Cancer Prevention Institute of California (formerly the Northern California Cancer Center) and the University of Melbourne (LeCalvez, 2011)) gathered women by population-based sampling methods, from which patients for our study were selected. Patients for these Breast CFRs were recruited between 1995 and 2005.

Patients designated as cases (N=1259) were selected based on diagnosis at or before the age of 45 years, self-reported race or ethnicity, and grandparents' country of origin which correlated with Caucasian, East Asian, Hispanic/Latino, or African American racial or ethnic heritage.

Control patients (N=1063) were frequency matched to case patients within each study center on racial or ethnic group, with age at selection for study not more than +/-

10 years difference of the age range at diagnosis of patients from the same center. A shortage of available controls in some age or ethnic groups caused the frequency matching to not be one-to-one in all subgroups.

SNP Genotyping

Genotyping began with whole-genome amplified (WGA) DNA. A nested polymerase chain reaction (PCR) with a multiplexed primary PCR was used. This was followed by high-resolution melting (HRM) curve analysis to identify major and minor alleles. Our panel of sixteen SNPs included variants in FGFR2, XRCC2, and the 8q24 genomic region, as shown in Table 1.

Our HRM analysis consisted of two assays, either unlabeled probe or small amplicon-based genotyping. Unlabeled probe was used whenever practical, and involved a probe of approximately 20 bases designed to be complimentary to the SNP of interest. Probes were designed to be complimentary to the major allele by convention. In cases where small amplicon-based genotyping was necessary, the secondary PCR amplicon was designed to be approximately 50 bases in length, so as to be able to visualize a single base change.

Genotyping began at the World Health Organization's (WHO) International Agency for Research on Cancer (IARC) before moving to the University of Utah's Huntsman Cancer Institute. The location where genotyping was done for each SNP is noted in Table 1.

Statistical Analysis

In order to calculate the polygene odds ratio for each subject, the appropriate SNP odds ratio as published by COGS was assigned to each genotype (Michailidou,

2013). Based on the 16 genotypes for each particular woman, the appropriate odds ratio values were then multiplied together to obtain the subject's crude polygene risk score. Furthermore, the average control polygene risk score was calculated. Each subject's crude polygene OR, case or control, was divided by the average control polygene risk score, to produce a normalized polygene risk score.

Using these normalized polygene risk scores (PRS), individual subjects were sorted into one of seven groups: PRS <0.5, 0.5 < PRS < 0.63, 0.63 < PRS < 0.8, 0.8 < PRS < 1.26, 1.26 < PRS < 1. 59, 1.59 < PRS < 2.0, PRS >2.0 (Table 2). Using the individual normalized polygene PRS in each group, an empirical estimate of the OR for each group was estimated (Table 2).

Table 1. Table of SNPs genotyped, rs numbers and major/minor allele, research center where subjects were genotyped for that SNP, the inheritance model (PA=per allele, REC=recessive), Odds Ratio (OR) as published by COGS, allele frequency (q) of the minor allele in the BCAC study as published by COGS, our observed OR, and our observed allele frequency (q).

Locus	SNP	Genotyped at	Model	iCOGS OR	iCOGS BCAC q	Observed OR	Observed q
FGFR2	rs2981579, G/A	IARC	PA	1.27	40%	1.15	48%
TOX3	rs3803662, G/A	IARC	PA	1.24	26%	1.4	27%
XRCC2	rs3218408	HCI	Rec	1.33		1.15	25%
5p12	rs10941679, A/G	HCI	PA	1.13	25%	1.02	28%
MAP3K1	rs889312, A/C	IARC	PA	1.12	28%	1.31	27%
SLC4A7	rs4973768, C/T	HCI	PA	1.10	47%	1.09	48%
8q24	rs13281615, A/G	IARC	PA	1.09	41%	1.13	41%
1p11.2	rs11249433, A/G	HCI	PA	1.09	40%	1.13	53%
ESR1	rs2046210, G/A	HCI	PA	1.08	34%	0.97	39%
ZMIZ1	rs704010, C/T	HCI	PA	1.08	39%	1.01	39%
LSP1	rs3817198, T/C	IARC	PA	1.07	31%	1.15	32%
ANKRD16	rs2380205, C/T	HCI	PA	0.98	44%	0.8	41%
CASP8	rs1045485, G/C	IARC	PA	0.97	13%	1.05	12%
COX11	rs6504950, G/A	HCI	PA	0.94	28%	0.72	33%
2q35	rs13387042, A/G	IARC	PA	0.88	49%	0.92	48%
ZNF365	rs10995190, G/A	HCI	PA	0.86	16%	0.83	20%

Table 2. Table of empirical Odds Ratio (OR). A crude PRS was calculated for each sample, and was then adjusted based on the average of the controls PRSs. Samples were then sorted into one of the normalized PRS groups listed and an OR for each group was calculated.

Normalized PRS Group	OR
OR <0.5	0.28
0.5<OR<0.63	0.62
0.63<OR<0.8	0.57
0.8<OR<1.26	1.00
1.26<OR<1.59	1.54
1.59<OR<2.00	1.29
OR >2.00	2.49

RESULTS

The distribution of subjects among our seven groups showed a bell curve centered on the reference group which straddles a normalized PRS of 1.0 (Figure 1). We found that the groups with a PRS <1.0 had a larger number of controls than cases, and conversely, those groups with a PRS >1.0 had a larger number of cases than controls. Thus, the OR for each group with an expected PRS <1.0 was found to be <1.0; similarly, those groups with a PRS >1.0 were found to have an OR >1.0 (Figure 1).

Our null hypothesis states that there would be no risk attributable to the polygene model, and so under the null hypothesis, on a graph of PRS vs. ORs (Figure 2), we would expect no association, with each OR being not significantly different than 1.0 (i.e. clustered around the y=1 line). Under the alternate hypothesis, that there is measurable risk attributable to the polygene, our groups should reflect the normalized product of the ORs from the original publications, and cluster around the line x=y. In fact we found the latter to be true (Figure 2); the slope of the regression line was 0.97 with a p-value=6.6*10-5, which overwhelmingly rejects the null hypothesis. We recognized that the large number of individuals in our central reference group, which by definition has an OR of 1.0, may have been acting to anchor our results, so in order to further challenge our results, we excluded the individuals in that central group and recalculated the slope and p-value. The recalculated slope was 1.04 and the p-value= 1.4*10-4 which still rejects the null hypothesis (Figure 3).

In looking at the possible clinical utility of this study, we next turned our attention to the portion of subjects whose normalized PRS reached the medically actionable threshold of equivalence to 20% increased risk (OR>2.0). We found that within our sample, 1.2% of controls and 3.6% of cases fell into that group. This percentage is similar to the number of individuals found to be at a medically actionable level of risk from single intermediate risk gene variants in recent studies by our and other labs (Tavtigian, 2009; Le Calvez-Kelm, 2011; Park, 2012).

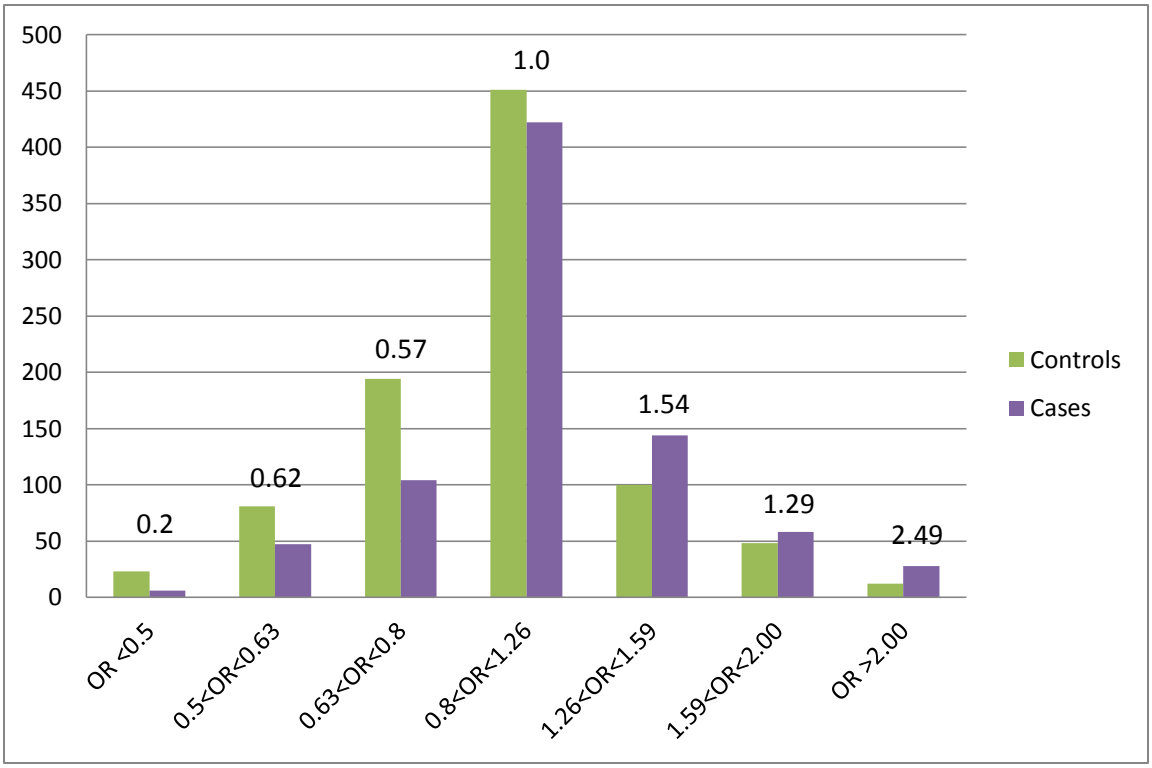

Figure 1. Graph of the number of Caucasian of European descent (CEU) cases and controls by normalized polygene risk score (PRS). The curve for controls is skewed toward ORs less than 1.0; the curve for cases is skewed toward ORs greater than 1.0. Number above bars equals the calculated empirical OR for each group.

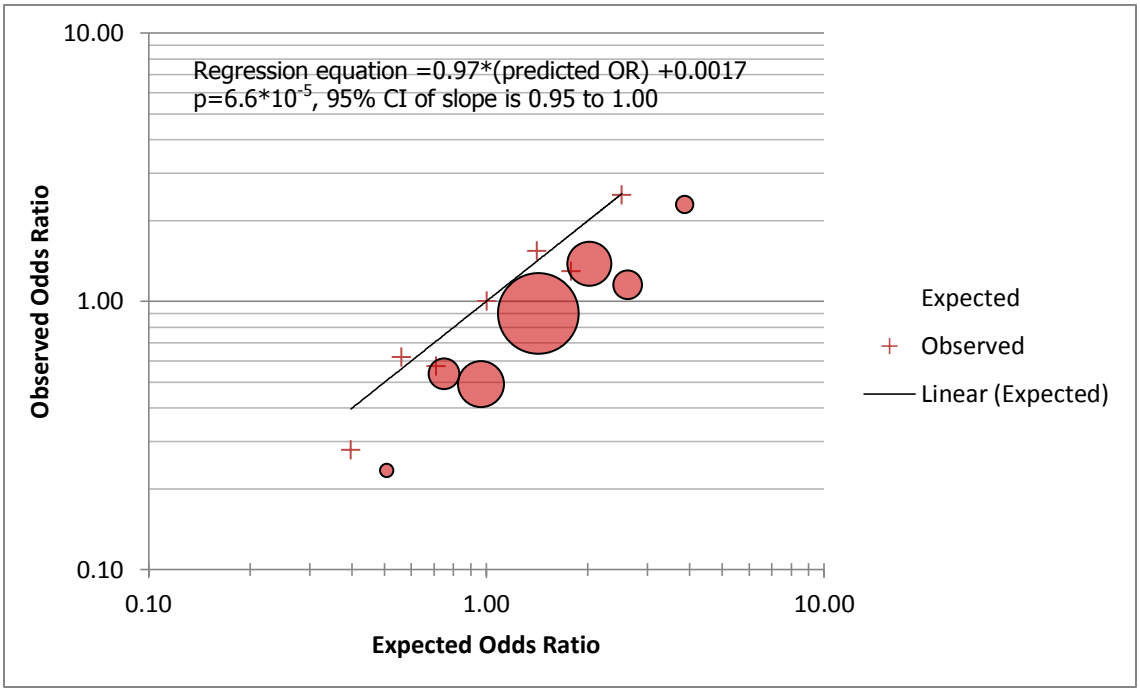

Figure 2. Graph of the expected vs. observed OR values by predicted odds ratio group. Red crosses are the empirical OR as presented in Table 2 and Figure 1. Each circle represents one of the OR groups with the bottom left circle representing OR<0.5 and the top right circle representing OR>2.0. The area of each circle is proportional to the number of samples in each group. The p-value of $6.6*10^{-5}$ strongly rejects the null hypothesis that there is no risk associated with the polygene.

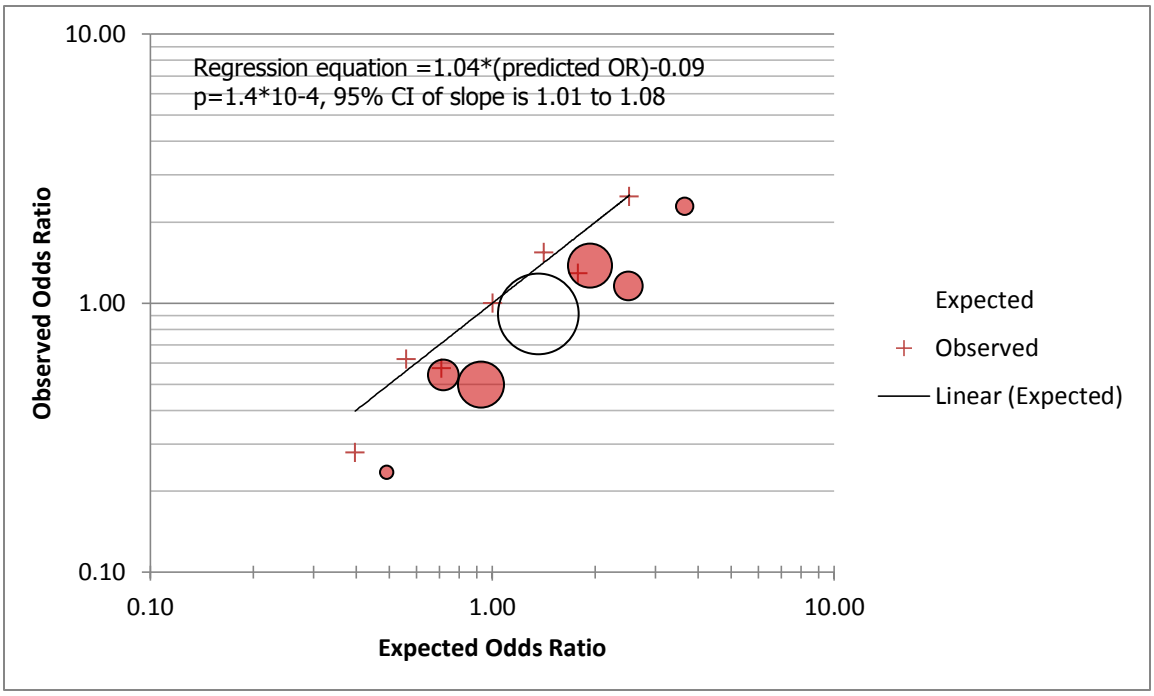

Figure 3. Graph of the expected vs. observed OR values by group when the central group is omitted from calculations. In order to determine whether the large central group was anchoring the observed regression, we recalculated the regression and p-value without the central group. The p-value of $1.4*10^{-4}$ still overwhelmingly rejects the null hypothesis.

DISCUSSION

Subjects

For our calculations we used previously published ORs as published by COGS and BCAC (Easton, 2007; Michailidou, 2013). Our study included a number of non-Caucasian subjects. Because the COGS and BCAC studies focused on Caucasian individuals of European descent, the allele frequencies and ORs are not necessarily directly applicable to subjects from other populations or ethnicities. Accordingly, we excluded data from non-Caucasian subjects from the calculations for odds ratio distribution.

Data Anomalies

When comparing our observed ORs and allele frequencies, we found a few notable anomalies when compared with the most recent COGS publication (Michailidou, 2013). In the case of SNP (rs11249433) 1p11.2, the published allele frequency is 40%, but in our sample, we observed the allele frequency to be 53%. In the case of SNP (rs2046210) ESR1 and SNP (rs704010) ZMIZ1, the published ORs were each just over 1.0, while we found their ORs to be slightly under 1.0. In the exact opposite case, SNP (rs1045485) Casp8 has a published OR of just under 1.0, while we observed the OR to be just over 1.0. In all cases, the discrepancies are likely due to our much smaller sample size when compared to the published COGS study. Additionally, because all of these ORs are very near to 1.0, they had a negligible effect on our composite PRS.

Common Disease/Common Variant vs. Common Disease/Rare Variant

The dichotomous models, Common Disease/ Common Variant (CD/CV) and Common Disease/Rare Variant (CD/RV), are often presented in a way that implies that one model or the other must be true, and that there is little room for an intermediate result or for the two models to work in concert. In fact, the data from this SNP genotyping study combined with case-control mutation screening of intermediate risk genes in these same subjects (Tavtigian, 2009; LeCalvez-Kelm 2011, Park 2012; unpublished Tavtigian lab work) seem to suggest that the CD/CV vs. CD/RV contrast is a false dichotomy. In fact, it seems that both models contribute to breast cancer susceptibility in the general population. Some individuals have notably elevated risk due to the common SNP polygene alone; others are likely to have elevated risk due to a single rare variant in an intermediate-risk or high-risk gene. Either way, obtaining a more complete estimate of an individual's genetically determined risk will require combining data from their common SNP genotypes with intermediate-risk and high-risk gene mutation screening data.

REFERENCES

Chakravarti, A. (1999) *Population genetics—making sense out of sequence.* Nat. Genet., 21: 56–60.

Chakravarti, A. (2001) *Single nucleotide polymorphisms: . . .to a future of genetic medicine.* Nature. 2001, 409: 822-823.

Easton, DF, et al. (2007) *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature. 2007, 447: 1087-1093.

Easton, DF and RA Eeles (2008) *Genome-wide association studies in cancer.* Hum. Mol. Genet., 2008, 17 (R2): R109-R115.

Ellsworth, R, et al. (2010) *Breast cancer in the personal genomics era.* Curr Genomics, 2010, 11(3): 146-161.

Harris, L, et al. (2007) *American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer*. J Clin Oncol. 2007, 25(33):5287-312.

Houlston, RS and J Peto (2004) *The search for low-penetrance cancer susceptibility alleles.* Oncogene, 2004, 23(38): 6471-6.

Howlader N, et al. (2009). *SEER Cancer Statistics Review, 1975–2009 (Vintage 2009 Populations)*, National Cancer Institute. Bethesda, MD, 2012.

Ivengar, SK and RC Elston (2007) *The genetic basis of complex traits: rare variants or "common gene, common disease"?* Methods Mol. Biol., 2007, 376: 71-84.

Lander, E.S. (1996) *The new genomics: global views of biology*. Science, 274: 536–539.

Le Calvez-Kelm, F, et al. (2011) *Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study.* Breast Cancer Research, 2011, 13:R6.

Michailidou, K, et al. (2013) *Large-scale genotyping identifies 41 new loci associated with breast cancer risk.* Nat Genet., 2013, 45(4): 353-61.

Park, DJ, et al. (2012) *Rare mutations in XRCC2 increase the risk of breast cancer.* Am J Hum Genet. 2012 Apr 6; 90 (4):734-9.

Pharoah, P. D. P. et al. (2002) *Polygenic susceptibility to breast cancer and implications for prevention.* Nature Genet., 2002, 31: 33–36.

Pritchard, JK (2001) *Are rare variants responsible for susceptibility to complex disease?* Am. J Hum. Genet., 2001, 69: 124-137.

Pritchard, JK and N Cox (2001) *The allelic architecture of human disease genes: common disease – common variant…or not?* Human Molecular Genetics. 2002: 11(20).

Reich, D.E. and Lander, E.S. (2001) *On the allelic spectrum of human disease*. Trends Genet., 17: 502–510.

Schork, NJ (2009) *Common vs. rare allele hypotheses for complex diseases.* Curr.Opin. Genet. Dev., 2009, 19(3): 212-9.

Tavtigian, SV, et al. (2009) *Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer.* Am J Hum Genet. 2009 Oct; 85 (4):427-46.

Wacholder, S, et al. (2010) *Performance of common genetic variants in breast-cancer risk models*. N Engl J Med., 2010, 362: 986-993.